DOI: https://doi.org/10.4038/sljas.v26i2.8201

Sri Lankan Journal of Applied Statistics, Vol (26-2)

A Novel Weighted Hybrid Method for Multiple Hypothesis Testing of Genomic Data

S. Ouchithya[1,*], N. Hettiarachchi[2], G. Dharmarathne[1], and D. Attygalle[1]

[1]University of Colombo, Department of Statistics, Colombo, Sri Lanka

[2]Tokyo Metropolitan Institute of Medical Science, Tokyo, Japan

*Corresponding author: susara.ouchithya@gmail.com

Received: 05th July, 2025/ Revised: 17th August, 2025/ Published: 25th August, 2025

ABSTRACT

Multiple Hypothesis Testing presents challenges due to increased false discoveries when conducting statistical tests simultaneously. Despite the proposal of novel correction techniques, the procedure of selecting the most suitable method has remained a black box. The trade-off that arises from controlling false positives and negatives through different correction techniques underlines the need for a cohesive framework. This study addresses the above challenges with a special focus on gene expression data and evaluates six widely used Multiple Hypothesis Testing (MHT) methods, namely, Bonferroni, Holm, sequential goodness of fit (SGoF), Benjamini-Hochberg, Benjamini-Yekutieli, and Storey's Q-values, across different scenarios to compare two independent groups. Our results show that Storey's Q-value performs well with large effect sizes, whereas SGoF excels in low-effect scenarios. However, the Bonferroni and Holm methods offer high precision owing to the strict control of false positives. Recognizing the limitations of relying on a single method, we introduce a novel Weighted Hybrid Method (WHM), a decision-support framework that allows users to navigate between approaches rather than serving as a new statistical test. An innovative Significant Index Plot (SIP) is unveiled to assist in the detection of significant hypotheses across different methods. The framework was tested on four genomic datasets: gene expression in multiple sclerosis (GSE21942), myelodysplastic syndrome (GSE61853), alcohol-related gene expression (GSE52553), and age-related corneal transcriptomes (GSE58315), extending the usage to enable independent hypothesis weighting. A novel Python library, MultiDST, and a web interface were

developed to enable researchers to apply the framework efficiently, improving the transparency of their findings.
Keywords: Multiple Hypothesis Testing; Significant Index plot; Gene Expression

1 Introduction

Statistical tests are essential in scientific research, enabling the comparison of alternative hypotheses by either rejecting or failing to reject a null hypothesis at a predetermined significance level. Each statistical test provides a p-value, which represents the probability of obtaining results as extreme, or more extreme than the ones observed, under the assumption that the null hypothesis is true. Although p-values have traditionally been the gold standard for evaluating significance, they fail to appropriately represent research findings under conditions such as multiple hypotheses (Ioannidis, 2005). A typical hypothesis test introduces two types of errors: Type I error, which involves incorrectly rejecting a true null hypothesis (false positive), and Type II error, which accepts a null hypothesis when the alternative is true (false negative). While traditional statistical tests focus primarily on controlling type I errors, challenges arise when dealing with multiple hypotheses simultaneously, leading to an increased risk of false discoveries. According to Diaconis (1985), if enough statistics are computed, some of them will be sure to show the structure. This phenomenon of "p-hacking" or selective inference poses challenges in the face of multiple testing without appropriate adjustments, resulting in a spurious excess of statistically significant results (Tukey, 1977). To overcome the above, various techniques have been developed, which are generally categorized into Family-Wise Error Rate (FWER) control methods and False Discovery Rate (FDR) control methods. Commonly used FWER control methods include Bonferroni correction, Šidák correction, and the Holm-Bonferroni method. Moreover, Benjamini-Hochberg and Storey's Q values are among the major FDR control methods. The issue of multiple comparisons is particularly pronounced in fields such as bioinformatics, economics, life science, and spatiotemporal analysis, where inferences are based on many statistical tests conducted simultaneously (Goeman and Solari, 2014; Cortés et al., 2020; Menyhart et al., 2021).

While controlling Type I errors helps limit false discoveries, it can also increase the likelihood of committing Type II errors (false negatives). Hence, strict Type I error control methods suffer from low statistical power (Goeman and Solari, 2014). This delicate balance, often referred to as a double-edged sword, is particularly problematic in fields such as genomics and medical research, where false negatives can be detrimental (Menyhart et al., 2021). One of the most common applications of multiple hypothesis testing (MHT) in genomics is evident in DNA microarrays, resulting in an array of p-values corresponding to the genes tested. Risch and Merikangas (1996) suggested

employing an extremely conservative p-value threshold of 5×10^{-8} to minimize false positives in the study of complex diseases. While this approach reduces false positives, it risks increasing false negatives due to stringency (Chen et al., 2021). Therefore, in genome-wide studies, it is essential to infer near-perfect statistical significance (Storey and Tibshirani, 2003). Although the issue of multiple comparisons has been debated for nearly a century, there is no definitive solution to obtain optimal results across all MHT scenarios. The selection of the best method is often subjective, as different correction techniques may provide varying results. Justifying the results here can be tedious, especially if the researcher is inexperienced with the statistical concepts surrounding the procedures (Menyhart et al., 2021).

To address this issue, our study seeks to analyse and compare the performance of six multiple testing methods through simulation studies, allowing for a novel iterative hybrid method as well as a weighting approach. The novel framework and visualizations are made available through a user-friendly Python library and a graphical user interface (GUI) that facilitates the implementation and comparison of various MHT methods. This library and GUI enable efficient identification of significant hypotheses from a given set of p-values. By unifying these methodologies into a comprehensive statistical framework, this study bridges a critical gap in MHT, empowering researchers towards reliable and robust results.

2 Methods

2.1 MHT correction methods considered in the study

Through an extensive examination and literature review, six most widely used candidates for MHT in genomic data analysis were selected, as shown in Table 1 below.

In addition, acknowledging that the general trend driven by the covariates (strata) may be the same across the different folds, independent hypotheses weighting (IHW) extensions were also considered. According to Ignatiadis et al. (2016), IHW is applicable in occasions where the covariates are informative of power or prior probability, independent of the p-values under the null hypothesis, and not notably related to the dependence structure of the joint statistics. If strong correlations among the tests are anticipated, it is important to select a covariate that does not significantly contribute to these correlations. In a differential gene expression study, a possible covariate is the sum of read counts per gene across all samples, and for GWAS, the minor allele frequency has been considered a candidate covariate (Love et al., 2014). Therefore, we consider applicable weights depending on the dataset.

Table 1: Correction methods selected for the study

Method	Reason for Selection	Error Type
Bonferroni Correction (Bonferroni, 1935)	Baseline method; simple and conservative	FWER
Holm-Bonferroni Correction (Holm, 1979)	Sequential approach; adjusts thresholds progressively	FWER
Benjamini-Hochberg Procedure (Benjamini and Hochberg, 1995)	Balances power and control; widely used in genomics	FDR
Benjamini-Yekutieli Method (Benjamini and Yekutieli, 2001)	FDR control under dependence; accounts for correlated tests	FDR
Storey's Q-Value (Storey and Tibshirani, 2003)	Adaptive FDR control; estimates proportion of true null hypotheses	FDR
SGoF Test (Sequential Goodness-of-Fit) (Carvajal-Rodríguez et al., 2009)	Increased power for large-scale testing; combines p-values from multiple tests	Varies (Depends on setup)

2.2 Metrics Used to Evaluate MHT Methods

The performance of each multiple correction method was evaluated using the following well established statistical metrics.

1. **Statistical Power:** Power measures the probability of correctly rejecting a false null hypothesis (True Positive Rate) given by $\frac{TP}{TP+FN}$. A method with higher power is preferable as it can detect more true positives.

2. **False Discovery Rate (FDR):** Measure the proportion of false positives among the rejected hypotheses calculated by $\frac{FP}{TP+FP}$.

3. **Balanced Accuracy:** Accuracy measures the proportion of correct predictions (both TP and TN) out of the total number of predictions calculated using $\frac{TP+TN}{TP+FP+TN+FN}$.

4. **Precision:** Precision or true positive rate (TPR) measures the proportion of predicted positives that are true positives given by $\frac{TP}{TP+FP}$. This is equivalent to $1 -$ FDR.

5. **Specificity:** Specificity or true negative rate (TNR) measures the ability to detect true negatives using $\frac{TN}{TN+FP}$.

6. **F1-score:** F1 score is the harmonic mean of precision ($\frac{TP}{TP+FP}$) and recall ($\frac{TP}{TP+FN}$), providing a balanced measure, which is particularly useful

when there is an imbalance between positive and negative instances. F1 score is given by $\frac{2\times\text{precision}\times\text{recall}}{\text{precision}+\text{recall}}$.

2.3 Bridging the Gap in Visualization with the Significance Index Plot

To overcome the lack of graphical visualization related to hypothesis rejection by different MHT correction methods, we designed an innovative custom plot, which will be referred to as the "Significant Index Plot (SIP)" from now on, by extending the "matplotlib" library available in Python (`https://matplotlib.org/`). This plot can be used to observe hypotheses rejected by each method, facilitating a thorough comparison of the performance of different methods.

When interpreting a SIP, each vertical line in front of each method given in the y-axis represents a significant hypothesis, while a white line (gap) corresponds to the lack of a signal. If a line extends across all methods from top to bottom, it means that the hypothesis was identified as significant by all the methods. For the simulation study, where true positives were known, the plot was modified to include an additional row for the true signals. This helps compare the performance of the different metrics more easily.

2.4 Design of the comparative simulation study

To evaluate the suitability of the selected methods, a comparative simulation was conducted employing a factorial design to systematically investigate the performance of the six MHT correction methods under two versions of independent samples t-tests: Student's test and Welch's t-test. The simulation parameters are varied across sample sizes, sample standard deviations, effect sizes, and true null proportions. All techniques here are implemented aiming at total FWER or FDR control under a level of 0.05, depending on the procedure, which can be extended to baseline thresholds such as 0.01 or 0.001, to detect signals based on the objectives of the study. Sample sizes are manipulated to values of 5, 15, and 30 to reflect different levels of statistical power, inspired by the experimental design of Kang et al. (2009). For the hypothesis tests conducted to obtain p-values, the effect sizes are varied to represent the magnitude of the true underlying effects. The effect sizes are set to 0.5, 1.0 and 1.5, corresponding to low, moderate and high effects, respectively, following the design of Carvajal-Rodríguez et al. (2009). Additionally, to assess the impact of sample variability of the groups, the sample standard deviations are varied under 0.5 and 1.0 by observing the behaviour of the p-values under both conditions. The true null proportion depicts the truly null hypotheses among all tested hypotheses, reflecting the prevalence of non-significant findings in the data. The metrics for the proportion of true nulls are set to 0.5, 0.75 and 0.9, corresponding to 50%, 25%, and 10% of the hypotheses being actual signals, respectively. These values were influenced by the simulation studies presented by Li et al. (2017). For each combination of parameter values,

100 Monte Carlo simulations were generated, influenced by the requirements behind the study, as well as the computational limitations.

2.5 Generating Simulated Test Data

Datasets of p-values were simulated via the same procedure presented earlier. A common seed value was used to generate distributions to allow for reproducibility. The simulations considered effect sizes of 0.5 and 1.5 and true null proportions (π_0 = 0.5, 0.8) to replicate realistic behavior. All the simulations were conducted under moderate sample sizes (n = 15) and high sample variability (S = 1.0) for ease of generalizability to gene expression scenarios.

2.6 Development of the Iterative Hybrid Framework

Our novel procedure stems from the observation that different methods may signal varying hypotheses owing to their inherent structure. Therefore, we present an exploratory hybrid technique to determine the significant hypotheses under most methods to sequentially test and remove until a satisfactory result is obtained, giving the choice to traverse both the FWER and FDR techniques, hence named "Hybrid". To execute the method, datasets of p-values were simulated, and a common seed value was used to generate distributions to allow for reproducibility. The simulations considered the following parameters (Table 2). Simulated experimental designs were considered under 04 different scenarios (high effect – high π_0, high effect – low π_0, low effect – high π_0, low effect – low π_0).

Table 2: p-value simulation parameters

Effect size	True null proportions (π_0)	Sample size (n)	Sample variability (S)
0.5 / 1.5	0.5 / 0.8	15	1.0

Figure 1 depicts the step-by-step procedure for conducting multiple hypothesis testing (MHT) with the methods discussed earlier. We suggest the following guidelines when applying the framework to make decisions regarding MHT scenarios.

1. **Visualize & validate** – Observe the histogram of the original distribution of p-values to obtain an idea of the potential effect size, true null proportion and sample variability.

2. **Apply classical MHT procedures** – Conduct a round of classical multiple hypothesis testing, identifying the hypotheses (genes) expressed by each of the methods. A significant index plot was used to gain a better understanding of the behaviour. This will be used as the baseline result.

3. **Incorporate weights (if available) & compare** – If individual weights are available, use the IHW procedure to obtain weighted p-values and

conduct a series of MHTs to explore the results given by the methods using the significant index plot. Compared with the baseline result.

4. **Iterative "Hybrid" procedure** – Note that the hypotheses that are marked as significant by the previously selected method are removed. If further improvement is needed, run additional iterations to select the method via continuous testing. We recommend visualizing the distribution and significant indices at each step to ensure effectiveness.

5. **Comparison of all methods** – Examine the results given by all the methods and choose the method that best suits the study and the underlying distribution. State-of-the-art methods are recommended for studies demonstrating higher effect sizes, whereas the hybrid methodology is more suitable when not enough signals are detected.

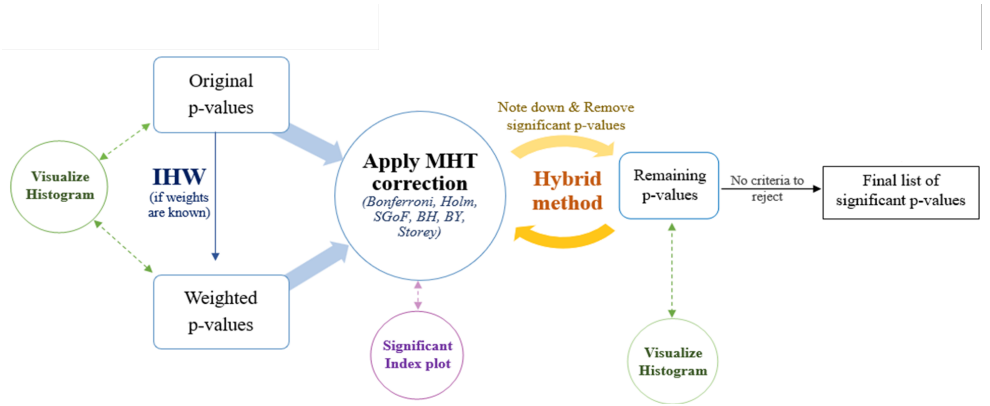

Fig. 1: Weighted Hybrid Method (WHM) framework

2.7 Application to Genomic Datasets

To evaluate the performance of our proposed weighted hybrid method (WHM) framework, we applied it to four publicly available gene expression datasets from the Gene Expression Omnibus (GEO) [`https://www.ncbi.nlm.nih.gov/geo/geo2r/`]. Datasets were chosen considering their relevance to the study of diverse biological conditions to assess the versatility and robustness of the framework. The novel framework was applied to both the raw p-values and the weighted values. As no weight information was available for the datasets, a reference was made to (Arsovski et al., 2015), who suggested that, compared with tandem duplicates, whole-genome duplicates tend to have twice the capacity for regulatory complexity. Therefore, in each of the datasets, the IHW scenario was carried out by allocating twice the weight for genes with known gene symbols (weight calculated as 1.025729) compared

with those with unknown symbols (weight calculated as 0.512865).

Dataset 01 – GSE 21942 (Multiple Sclerosis vs Healthy) contains gene expression data from the peripheral blood mononuclear cells (PBMCs) of multiple sclerosis (MS) patients and healthy controls, and identifies differentially expressed genes and pathways involved in MS pathogenesis. The dataset includes 27 samples, including 15 controls and 12 patients (Kemppinen et al., 2011).
`https://www.ncbi.nlm.nih.gov/geo/geo2r/?acc=GSE21942`

Dataset 02 – GSE 61853 (Myelodysplastic Syndrome (MDS) vs Control) contains gene expression profiles from bone marrow mesenchymal stromal cells (BM MSCs) of myelodysplastic syndrome (MDS) patients and healthy controls. This dataset consists of 14 samples, of which 7 are controls and 7 are from MDS patients (Kim et al., 2015).
`https://www.ncbi.nlm.nih.gov/geo/geo2r/?acc=GSE61853`

Dataset 03 – GSE 52553 (Alcoholic vs Non-Alcoholic) explores expression differences in lymphoblastoid cell lines (LCLs) from alcoholics and non-alcoholics after a 24-hour ethanol treatment. This study aimed to identify subtle gene expression changes caused by ethanol exposure. The dataset contains 42 samples, with 21 alcoholics and 21 controls (McClintick et al., 2014).
`https://www.ncbi.nlm.nih.gov/geo/geo2r/?acc=GSE52553`

Dataset 04 – GSE 58315 (Paediatric vs Adult) contains transcriptome profiles of human corneal endothelial cells (HCEnCs), aiming to characterize gene expression patterns across age groups for 6 samples from paediatric donors and 5 samples from adult donors, examining the potential connection of age groups to corneal endothelial dystrophies (Frausto et al., 2014).
`https://www.ncbi.nlm.nih.gov/geo/geo2r/?acc=GSE58315`

3 Results and Discussion

3.1 Results of the Comparative Study

The first simulation of the comparative study is focused on the equal variance t-test. When examining the statistical power over effect sizes, as shown in Figure 2, Storey's Q-value consistently exhibited the highest power, particularly with increasing sample size. Additionally, Benjamini-Hochberg (BH) and Benjamini-Yekutieli (BY) methods demonstrated strong performance, especially for larger effect sizes.

In contrast, the Bonferroni and Holm corrections showed limited power at smaller sample sizes (n = 5) but improved moderately with increasing sample size. The Holm method performed slightly better than the Bonferroni method did, whereas the SGoF procedure outperformed the other methods at n = 5 when low effect sizes were detected (0.5), demonstrating its ability to detect

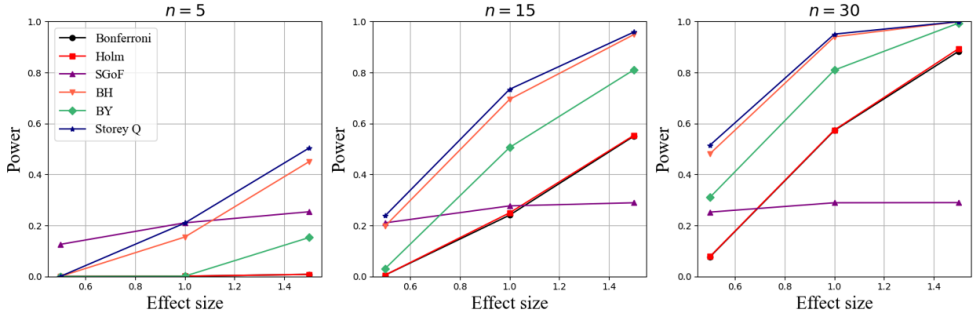

Fig. 2: Statistical Power over effect size grouped by sample size

weak effects more accurately.

When comparing power against the proportion of true nulls (π_0), all methods except SGoF exhibited an increase in power as the proportion of true signals (π_1) rose, as shown in Figure 3. The behaviour of the meta-test-based SGoF was distinct, which led to an inflation of false positives in the presence of a higher proportion of true signals. Across all sample sizes, Storey's Q value remained robust, outperforming the other methods. All methods benefitted from reduced variability in the samples, leading to higher power, as shown in Figure 4. Here, again, Storey's Q value stood out, with Benjamini-Hochberg and Benjamini-Yekutieli following closely.

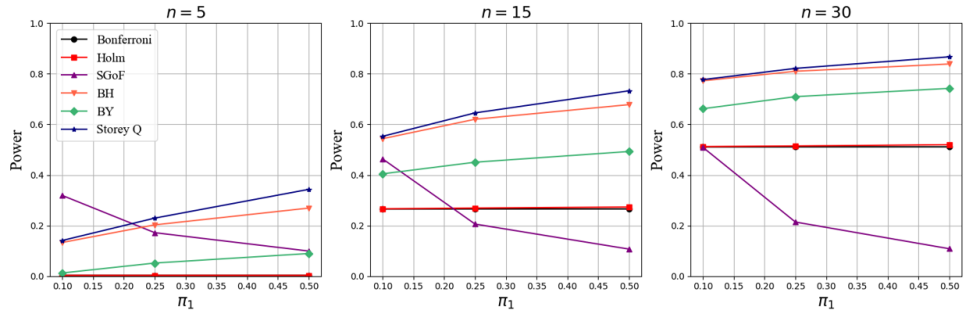

Fig. 3: Statistical Power over π_1 grouped by sample size

In the second simulation study involving Welch's t-test, the statistical power was highly influenced by sample size. As the size increased in at least one group, all methods displayed improvements in power, as shown in the heatmap in Figure 5. Storey's Q value and Benjamini-Hochberg stood out, showing consistently higher power at larger sample sizes (n = 15). The SGoF method showed better performance when the sample size was smaller (n = 5), maintaining its strength in small sample size settings. As shown in Figure 6, lower

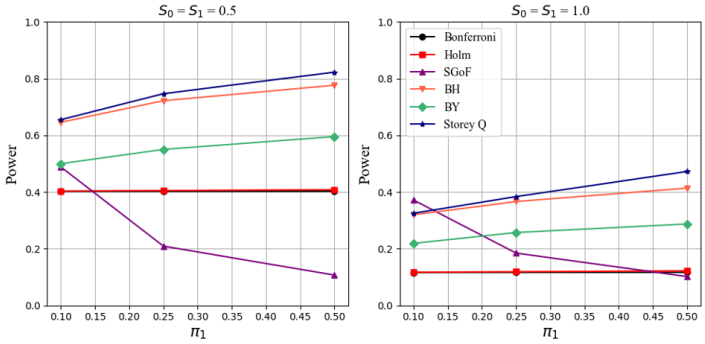

Fig. 4: Statistical Power over π_1 grouped by sample standard deviation

variability in either sample group led to greater power across all methods, with Storey's Q value and Benjamini-Hochberg once again leading.

The radar plot analysis in Figure 7 revealed that Storey's Q value was the top-performing method across all evaluation metrics, including power, specificity, precision, accuracy, and F1 score, especially in larger sample settings, whereas SGoF presented a higher F1 score when sample sizes were small (n = 5), making it a useful choice in cases involving a high proportion of true null hypotheses or imbalanced designs. Although the methods of Bonferroni and Holm demonstrated higher precision due to their conservative approach, they are most suitable in scenarios where strict control of false positives is crucial, despite their lower power.

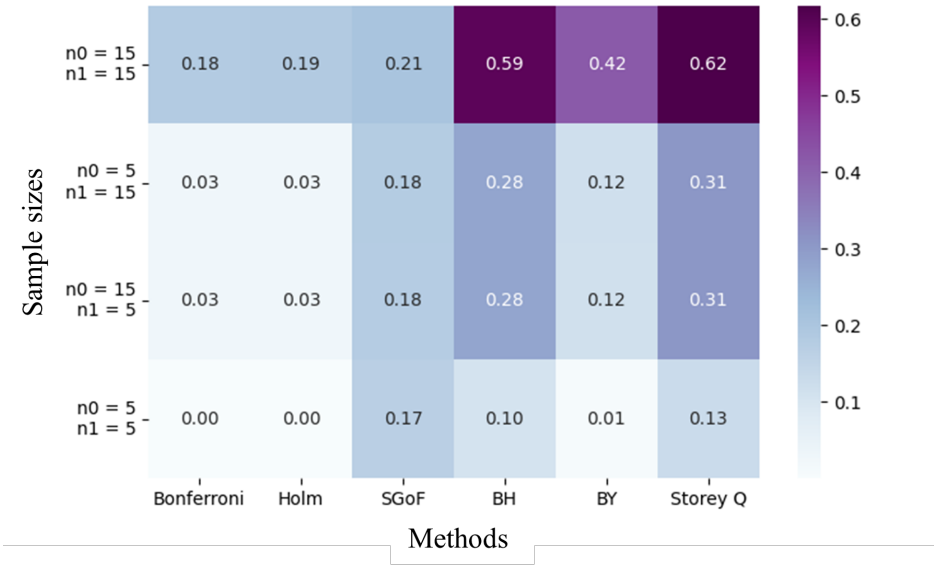

Fig. 5: Heatmap of power over sample size combinations

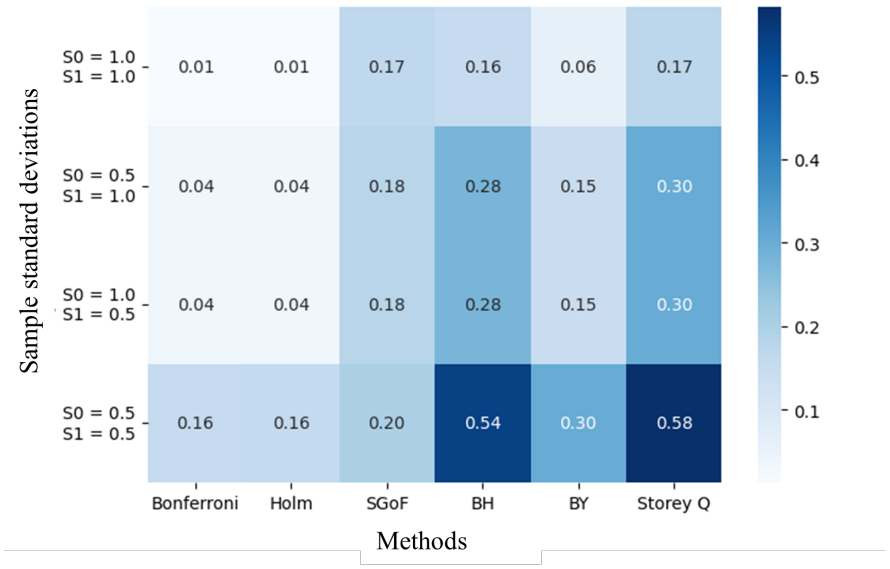

Fig. 6: Heatmap of power over sample standard deviation

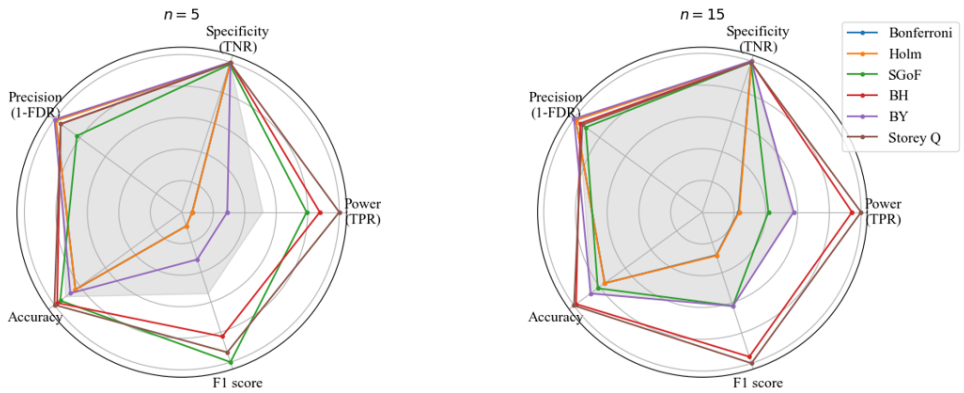

Fig. 7: Radar plot for comparing performance under Welch's test

Setting 01: High Effect, High π_0

This scenario (effect size = 1.5, π_0 = 0.8) exhibited the characteristic steep slope of a high true null proportion, as evident in Figure 8 (left). The Bonferroni and Holm methods initially detected 210 significant hypotheses each, whereas the SGoF method identified 543. Storey's Q-value showed the most promising results, with 1938 hypotheses detected at a power of 0.91. Upon removal of the p-values signalled by Storey's Q-value, most methods, except for SGoF, struggled to provide meaningful results in subsequent iterations. However, as in the SIP given in Figure 9, not all SGoF signals are true positives, hence leading to an increased risk of false positives, which is also suggested

by the histogram of p-values after the first removal shown in Figure 8 (right) due to the lack of a steep slope. Therefore, Storey's Q-value provides a reliable estimate of the true null proportion, validating its suitability for high effects and high π_0 settings.

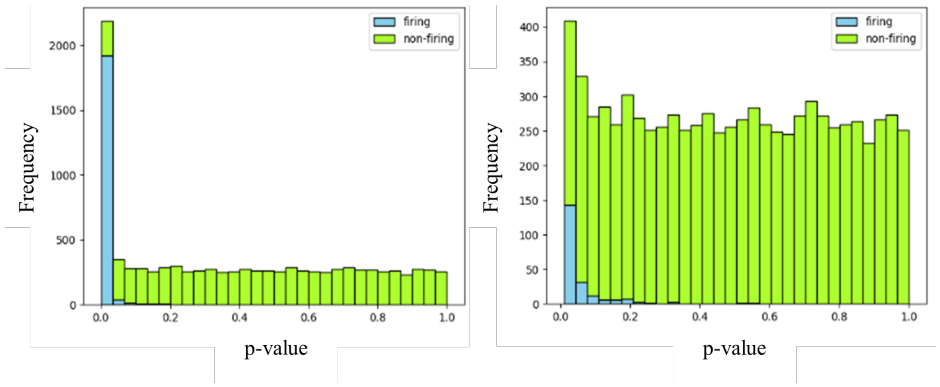

Fig. 8: Histogram of initial p-values and p-values after the removal of Storey's estimate

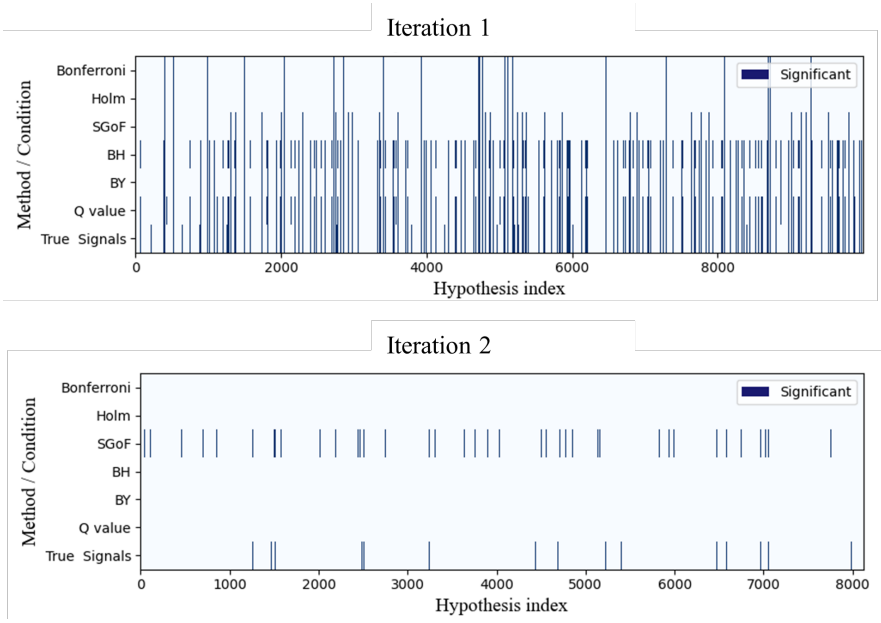

Fig. 9: Significant Index Plots for the two iterations

Setting 02: High Effect, Low π_0
For this simulation (effect size = 1.5, $\pi_0 = 0.5$), the histogram depicted a more pronounced cliff, reflecting both the high effect size and the lower proportion

of true null hypotheses, as shown in Figure 10. The hybrid procedure rejected 5142 hypotheses, with Storey's Q-value dominating other methods in terms of both power and the number of significant hypotheses detected (5142 out of 5000 firing hypotheses). No other method provided additional significant results after this iteration, affirming Storey's Q-value as the optimal choice for high-effect, low-π_0 scenarios (Figure 11).

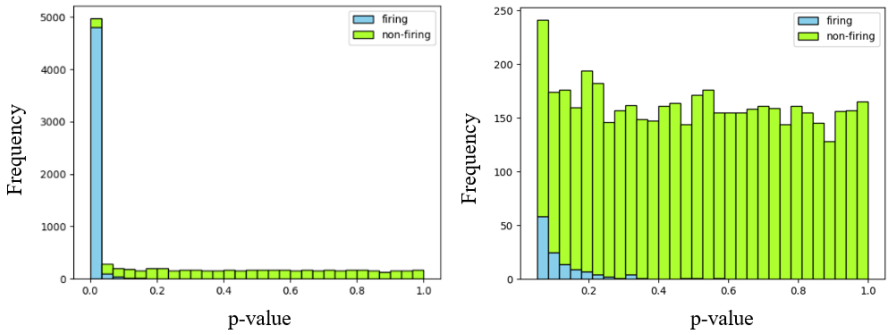

Fig. 10: Histogram of initial p-values and p-values after the removal of Storey's estimate

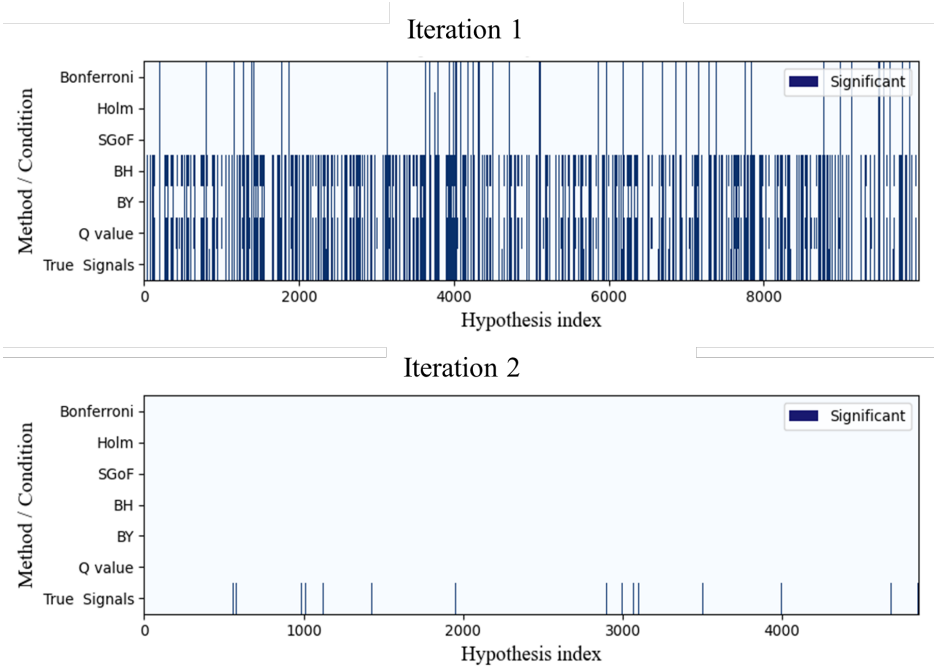

Fig. 11: Significant index plots for the two iterations in setting 02

Setting 03: Low Effect, High π_0

Under this configuration (effect size = 0.5, π_0 = 0.8), only SGoF identified 543

significant hypotheses. The pattern of the histogram of p-values is also less skewed than those of settings 01 and 02, as shown in Figure 12. However, the low effect size and high true null proportion led to an overestimation of π_0, causing a reduction in power. The SGoF proved to be effective under these challenging conditions, where detecting truly significant hypotheses was more difficult. However, as shown in Figure 13, not all true signals are detected, and there is also a tendency to pick up false positive results since the given power is only 0.243.

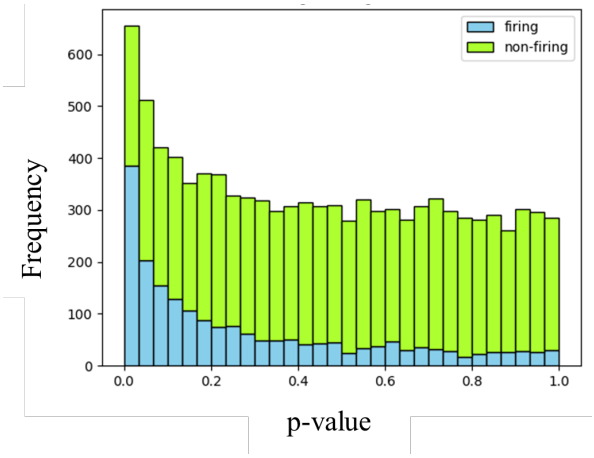

Fig. 12: Histogram of p-values under setting 03

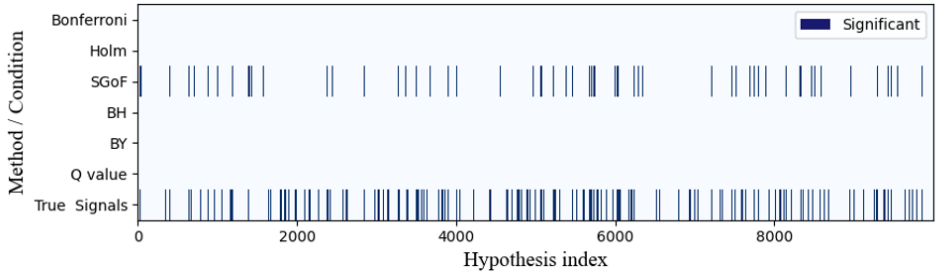

Fig. 13: Significant index plot for the first iteration under Setting 03

Setting 04: Low Effect, Low π_0
With effect size = 0.5 and π_0 = 0.5, the p-values in Figure 14 show a slightly higher skew than for setting 03, although it is less than that in high effect size scenarios. SGoF again provided the highest number of detections (543 significant hypotheses, 484 true positives), as shown in Figure 15. Storey's Q-value detected only 17 significant hypotheses. Over 25 iterations, the number of significant hypotheses detected by SGoF gradually decreases, suggesting that

while SGoF performs well initially, repetitive iterations lead to low signals.

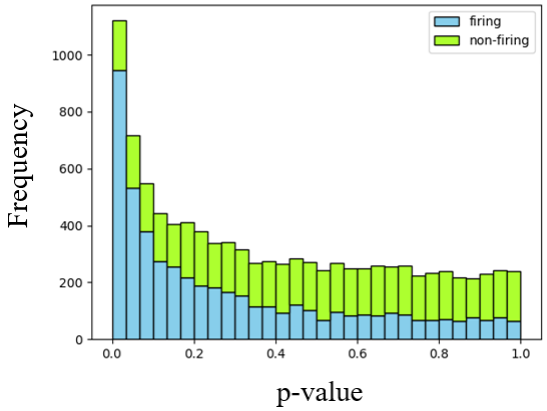

Fig. 14: Histogram of p-values under setting 04

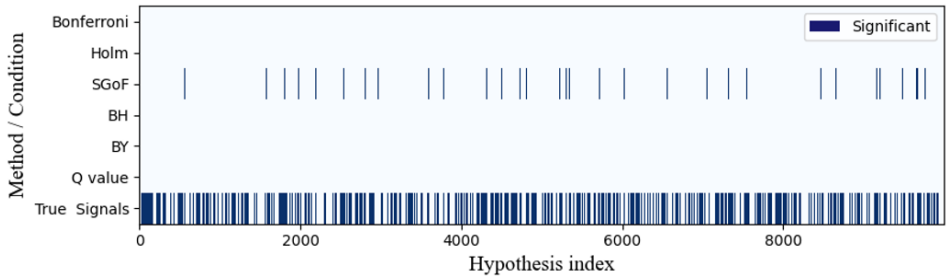

Fig. 15: Significant index plot for the first iteration under Setting 04

With different settings, visualization plays a crucial role in enhancing the iterative method's performance, making it essential to review plots before making decisions. Storey's Q-value excelled in scenarios with high effect sizes, whereas the SGoF was the only consistent performer with low effect sizes. A minor difference between the number of BH and Q-value detections suggested that Storey's procedure lacks significant conservativeness and is preferable when more hypotheses are detected.

3.2 Results from applying WHM to genomic datasets

Our novel weighted hybrid method (WHM) framework was extensively tested and verified on four publicly available gene expression datasets covering diverse biological conditions. Below, we provide the results for each dataset, focusing on the number of significant hypotheses detected under both unweighted and weighted procedures.

Dataset 01 – GSE 21942 (Multiple Sclerosis vs Healthy)

This dataset contained 54675 p-values, resembling a "Setting 01" simulation, with a high effect size and high true null proportion, as shown in Figure 16. The steep cliff-like nature is evident in the histograms of both scenarios before and after IHW. After applying the six correction techniques, Storey's Q-value proved to be the most suitable procedure for aligning with the observations of the simulation study. According to the SIP in Figure 17, Storey's Q-value and BH both show similar behaviour, suggesting fewer issues due to conservativeness by choosing Storey's Q-value.

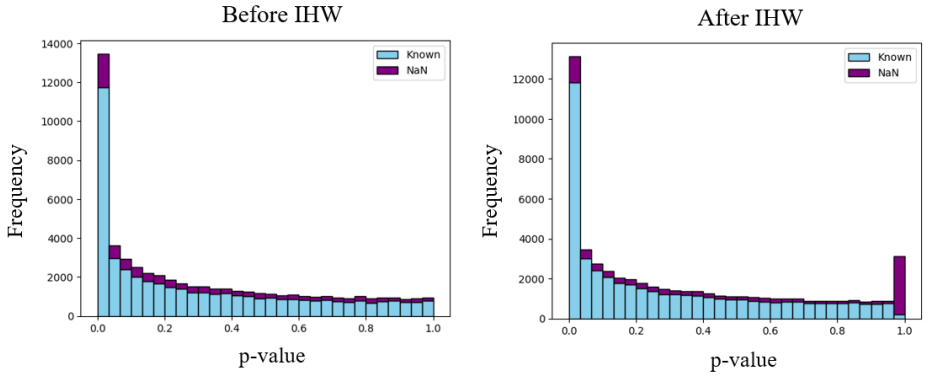

Fig. 16: Histogram of p-values for GSE 21942

As shown in Table 3, the WHM framework detected a total of 11767 and 9242 significant hypotheses using Storey's Q-value under the unweighted and weighted methods, respectively. This reduction reflects the ability of the refined weighting scheme to adjust for underlying biological variance.

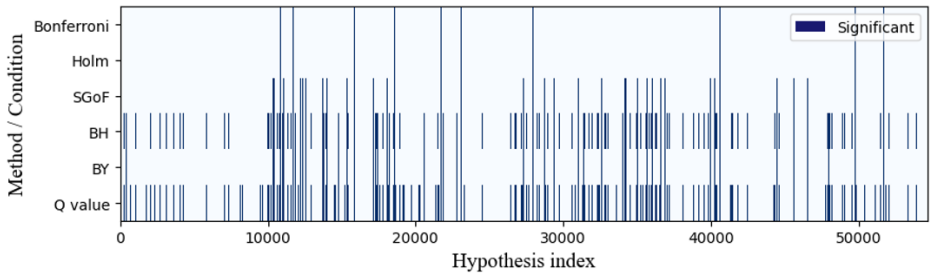

Fig. 17: Significant Index Plot for GSE 21942

Table 3: MHT Results for GSE 21942 (Number of hypotheses marked as significant)

Type	Uncorrected	Bonferroni	Holm	SGoF	BH	BY	Q-value
Before IHW	15406	893	898	2836	8904	4086	11767
After IHW	15012	884	890	2836	8686	4025	9242

Dataset 02 – GSE 61853 (MDS vs Control)
This dataset contained 47323 p-values. Figure 18 resembles a "Setting 03" simulation, suggesting a low effect size and a relatively high true null proportion. The SIP in Figure 19 depicts the same behaviour as that of the simulated scenario, with only SGoF showing noteworthy detections.

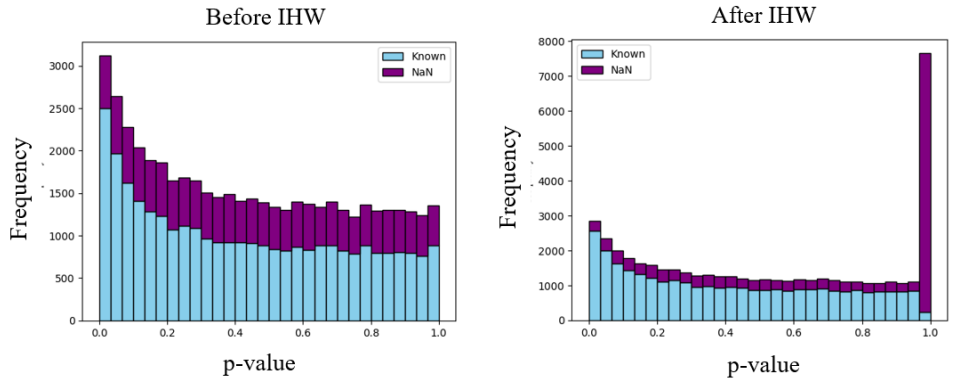

Fig. 18: Histogram of p-values for GSE 61853

As shown in Table 4, the WHM framework identified 2465 significant results with unweighted SGoF and 2461 with IHW. The Bonferroni, Holm, BH, and Q-values detected only one p-value of 3.75×10^{-7} with the gene ID *ILMN_1801776*. Owing to the limited number of low p-values, with only 4518

and 4071 p-values out of 47323 below the threshold of 0.05, it is inconclusive whether SGoF is the best method.

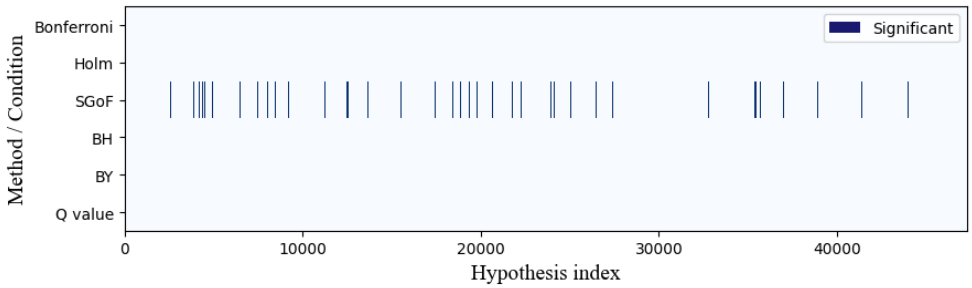

Fig. 19: Significant Index Plot for GSE 61853

Table 4: MHT Results of GSE 61853 (Number of hypotheses marked as significant)

Type	Uncorrected	Bonferroni	Holm	SGoF	BH	BY	Q-value
Before IHW	4518	1	1	2465	1	0	1
After IHW	4071	1	1	2461	1	0	1

Dataset 03 – GSE 52553 (Alcoholic vs Non-Alcoholic)
This dataset contained a total of 54675 p-values. Akin to Dataset 02, the histogram of p-values in Figure 20 depicts a "Setting 03" simulation but tends slightly towards "Setting 04" with a slightly steeper cliff, suggesting either a relatively greater effect size or a lower true null proportion. The observations in histograms are verified by the SIP given in Figure 21, with only the SGoF technique showing visible lines but more than for dataset 02.

Table 5: MHT Results for GSE 52553 (Number of hypotheses marked as significant)

Type	Uncorrected	Bonferroni	Holm	SGoF	BH	BY	Q-value
Before IHW	5060	5	5	2838	60	0	74
After IHW	4872	5	5	2840	62	0	62

Here, the WHM framework detected 2838 significant indices under the unweighted SGoF method and 2840 under the IHW method, as shown in Table 5. However, there are more detections in other methods in this case, where the Q-value also indicates 74 and 62 signals under the two approaches. The number of detections is also close to BH, suggesting fewer issues because it is

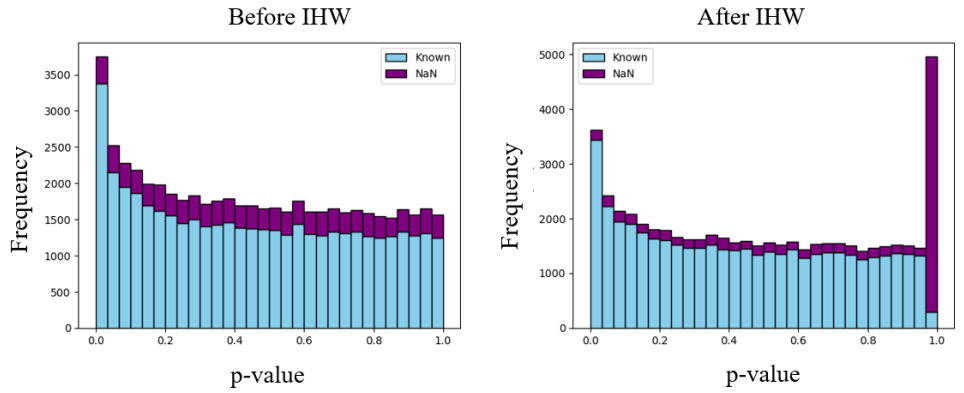

Fig. 20: Histogram of p-values for GSE 52553

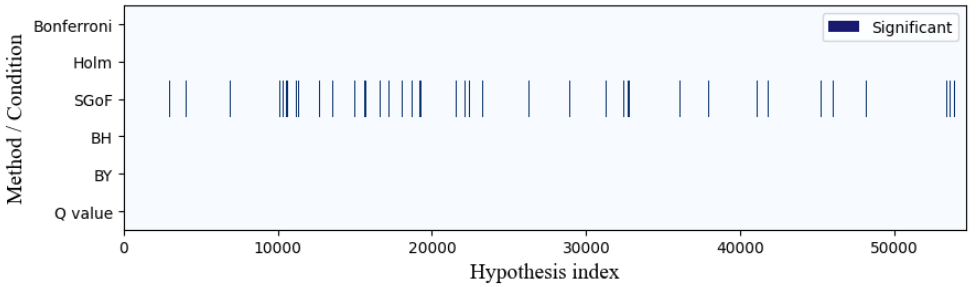

Fig. 21: Significant Index Plot for GSE 52553

less conservative. Furthermore, both Bonferroni and Holm methods detected 5 signals each. However, in this case, given that only 5060 and 4872 out of the total 54675 p-values were under 0.05, it is inconclusive to suggest a single method. Based on the level of accuracy needed, either the Q-value or SGoF can be chosen.

Dataset 04 – GSE 58315 (Paediatric vs Adult)

A total of 33297 p-values were extracted from GSE 58315. As Figure 22 depicts, the skewed nature is almost dissipated, and the histogram resembles the "Setting 03" or "Setting 04" simulation after the removal of Storey's detections. These observations are verified by the SIP in Figure 23, which shows signals only under SGoF and the SGoF detected p-values of Figure 24. The results after multiple testing given in Table 6 revealed that only SGoF had 1744 p-values. The inability of other methods to detect a single significant p-value compelled us to further analyse SGoF detected p-values, where SGoF signals were separately taken through the WHM procedure where a pattern such as the "Setting 04" simulation was obtained. However, because all the

p-values are now under the threshold of 0.05, the less conservative methods tend to hallucinate. This is detected by the SIP given in Figure 25. However, owing to the reduction in the total set of p-values, now previously more conservative methods, such as the Bonferroni and Holm methods, detected 5 genes each as significant.

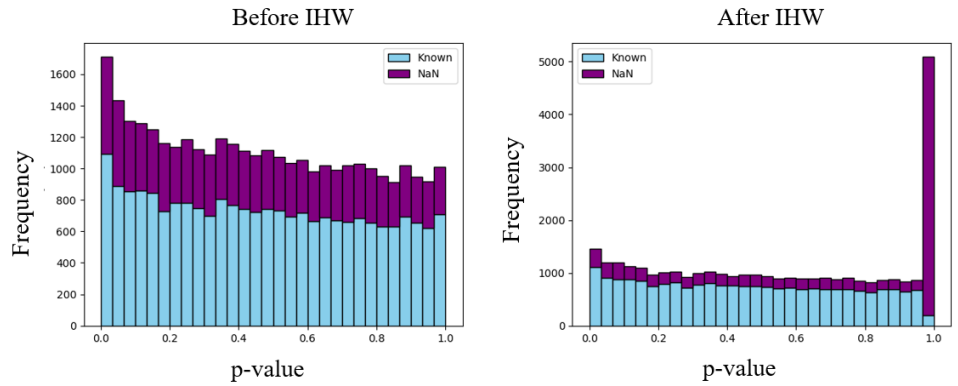

Fig. 22: Histogram of p-values for GSE 58315

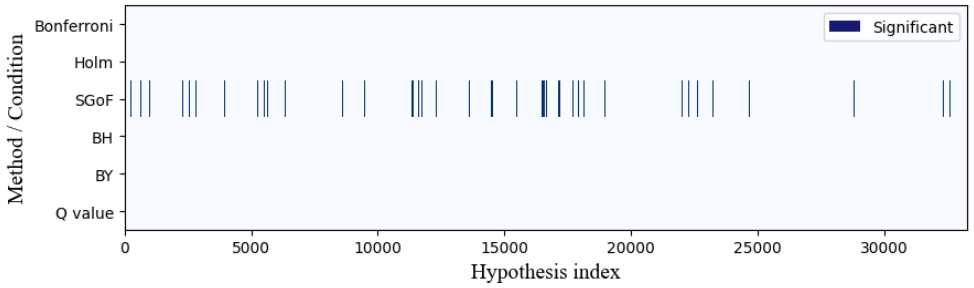

Fig. 23: Significant Index Plot for GSE 58315

Summary of Results

As evident from the results, different types of genomic datasets can be seamlessly integrated into the WHM framework to uncover insights, as summarized in Table 7 below.

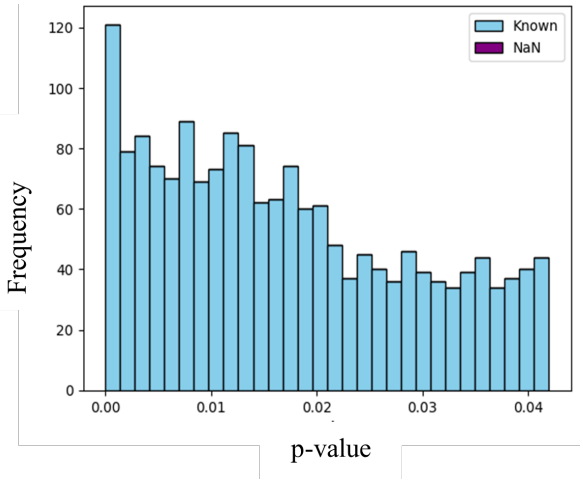

Fig. 24: Histogram of p-values for SGoF detected p-values

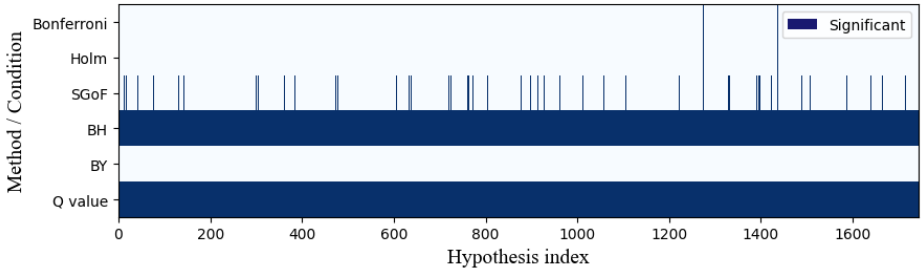

Fig. 25: Significant Index Plot for SGoF detected p-values

Table 6: MHT Results for GSE 58315 (Number of hypotheses marked as significant)

Type	Uncorrected	Bonferroni	Holm	SGoF	BH	BY	Q-value
Before IHW	2488	0	0	1744	0	0	0
After IHW	2086	0	0	1744	0	0	0
SGoF*	1744	5	5	105	1744	0	1744

Note: SGoF* refers to results obtained by reapplying MHT on the p-values initially detected by SGoF.

4 Conclusion

This study addresses the longstanding challenge of multiple hypothesis testing (MHT) in genomic data by developing a unified, user-friendly framework. We focused on optimizing six key MHT methods: Bonferroni, Holm-Bonferroni, Benjamini-Hochberg (BH), Benjamini-Yekutieli (BY), Storey's

Table 7: Summary of Results

Dataset	Resembling Simulation Setting	Suggested MHT correction	Total p-values	Unweighted signals	IHW signals
GSE 21942	Setting 01	Q-value	54,675	11,767	9,242
GSE 61853	Setting 03	SGoF / Bonferroni, Holm, BH	47,323	2,465	2,461
GSE 52553	Setting 03 (tending towards Setting 04)	SGoF / Q-value	54,675	2,838 (SGoF), 74 (Q-value)	2,840 (SGoF), 62 (Q-value)
GSE 58315	Setting 03 (Setting 04 for SGoF)	SGoF, then Bonferroni/ Holm	33,297 (1744)	1744 (SGoF), 5 (Bonferroni & Holm)	1744 (SGoF), 5 (Bonferroni & Holm)

Q-value, and Sequential Goodness of Fit test (SGoF). Extensive simulation studies were conducted across varying effect sizes and true null proportions, revealing significant insights into their performance. Storey's Q-value method demonstrated the highest power for larger effect sizes, whereas the SGoF method outperformed the other methods in settings of low effect sizes and high true null proportions, confirming the findings of Carvajal-Rodríguez et al. (2009) on the ability of the SGoF method to detect signals as the number of tests increases. However, with fewer tests, SGoF showed an increased risk of false positives, consistent with the caution laid by Castro-Conde and De Uña-álvarez (2014) on the reduced reliability of SGoF in small-sample scenarios.]

Following the comparative study, we introduced a novel hybrid iterative testing procedure that allows switching between FWER and FDR error control strategies as significant hypotheses are removed, enabling users to tailor the method to their specific datasets. This is particularly useful in gene-expression studies, where prior knowledge about certain genes can guide the testing strategy for more representative outcomes. Our novel hybrid method demystifies the "black box" nature surrounding multiple testing by providing a framework that can be extended to various practical scenarios that can be adapted for different datasets, enhancing its versatility to a range of MHT problems beyond genomic data. Rather than relying solely on the performance of a single correction technique, our approach integrates multiple methods to support the end user's decision-making, tailored to the specific experimental context. Furthermore, to address the limited visualization tools in MHT, we developed the "Significant Index Plot", enabling users to visually track signals identified

by different methods.

The framework was tested on four genomic datasets obtained from GEO2R, demonstrating its effectiveness. The findings highlight the ability of the WHM framework to handle various distributions of p-values while providing meaningful results. Dataset 01 (GSE 21942) resembled a "Setting 01" simulation, emphasizing the method's ability to handle datasets with strong signals. Moreover, in Datasets 02 (GSE 61853) and 03 (GSE 52553), where the p-value distributions resembled those of the "Simulation 03" scenario, the SGoF method outperformed the other methods, whereas the other methods detected minimal signals. However, in Dataset 04 (GSE 58315), none of the methods except SGoF managed to detect significant hypotheses. Here, the SGoF-detected p-values were separated and reanalysed to obtain more information.

The analysis of datasets demonstrates how the framework can be tailored based on the scientist's knowledge of the dataset and desired level of stringency, ensuring meaningful and reliable results. To ensure accessibility, we created the open-source Python package "MultiDST", which integrates all six methods, the hybrid approach, and visualization tools into a cohesive platform. The library is readily available on the Python Package Index (PyPI) at `https://pypi.org/project/multidst/` and can be installed by running the command "pip install multidst". Detailed documentation on how to use the Python library can be found at `https://multidst-docs-2.readthedocs.io/en/latest/multidst.html`. This package bridges the gap between statistical techniques and practical implementation in gene expression analysis, streamlining multiple testing corrections. In addition to the software package, a user-friendly website (GUI) was also developed (`https://multidst-frontend.vercel.app/`) to assist non-programmers in utilizing the framework.

In summary, our study successfully unified key MHT methods into a single framework focused on genomic data. The novel hybrid testing procedure and Significant Index Plot provide enhanced tools for analysing complex datasets. The adaptability of the WHM framework allows for extension to various practical scenarios, even beyond the scope of gene expression data. The framework can be easily applied to many other fields involving multiple testing, such as neuroscience, economics and computer science. Therefore, based on the simulation study and the application to real datasets, the framework demonstrated practical utility, and the release of the MultiDST Python package and the website aims to empower comprehensive, accurate analyses of data in scenarios involving multiple hypothesis testing.

Code Repositories

Python library:
`https://github.com/susara-ouchi/MultiDST_package`

Simulation study:
`https://github.com/susara-ouchi/MultiDST_simulations`

Analysis of genomic datasets:
`https://github.com/susara-ouchi/MultiDST-Analysis`

List of Abbreviations

BH Benjamini-Hochberg

BY Benjamini-Yekutieli

FDR False Discovery Rate

FWER Family Wise Error Rate

IHW Independent Hypothesis Weighting

MHT Multiple Hypothesis Testing

TNR True Negative Rate

TPR True Positive Rate

References

Arsovski, A.A. et al. (2015) Evolution of Cis-Regulatory Elements and Regulatory Networks in Duplicated Genes of Arabidopsis, *Plant Physiology*, 169(4), pp. 2982–2991.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, *Journal of the Royal Statistical Society: Series B (Methodological)*, pp. 289–300.

Benjamini, Y. and Yekutieli, D. (2001) The Control of the False Discovery Rate in Multiple Testing under Dependency, *The Annals of Statistics*, 29(4), pp. 1165–1188.

Bonferroni, C. (1935) Il calcolo delle assicurazioni su gruppi di teste, in.

Carvajal-Rodríguez, A., de Uña-Alvarez, J. and Rolán-Alvarez, E. (2009) A new multitest correction (SGoF) that increases its statistical power when increasing the number of tests, *BMC Bioinformatics*, 10, p. 209.

Castro-Conde, I. and De Uña-álvarez, J. (2014) sgof: An R package for multiple testing problems, *R Journal*, 6(2), pp. 96–113.

Chen, Z. et al. (2021) Revisiting the genome-wide significance threshold for common variant GWAS, *G3 Genes|Genomes|Genetics*, 11(2), p. jkaa056.

Cortés, J. et al. (2020) Accounting for multiple testing in the analysis of spatio-temporal environmental data, *Environmental and Ecological Statistics*, pp. 293–318.

Diaconis, P. (1985) Theories of Data Analysis: From Magical Thinking Through Classical Statistics, in, pp. 1–36.

Frausto, R.F., Wang, C. and Aldave, A.J. (2014) Transcriptome analysis of the human corneal endothelium, *Investigative Ophthalmology and Visual Science*, 55(12), pp. 7821–7830.

Goeman, J.J. and Solari, A. (2014) Tutorial in biostatistics: Multiple hypothesis testing in genomics, *Statistics in Medicine*, pp. 1–27.

Holm, S. (1979) A Simple Sequentially Rejective Multiple Test Procedure, *Scandinavian Journal of Statistics*, 6(6), pp. 65–70.

Ignatiadis, N. et al. (2016) Data-driven hypothesis weighting increases detection power in genome-scale multiple testing, *Nature Methods*, 13(7), pp. 577–580.

Ioannidis, J.P.A. (2005) Why Most Published Research Findings Are False?, *Evolution Psychiatrique*, pp. 443–454.

Kang, G. et al. (2009) Weighted multiple hypothesis testing procedures, *Statistical Applications in Genetics and Molecular Biology*, 8(1).

Kemppinen, A.K. et al. (2011) Systematic review of genome-wide expression studies in multiple sclerosis, *BMJ Open*, 1(1), p. e000053.

Kim, M. et al. (2015) Increased Expression of Interferon Signaling Genes in the Bone Marrow Microenvironment of Myelodysplastic Syndromes, *PLOS ONE*, 10(3), p. e0120602.

Li, D. et al. (2017) Bon-EV: An improved multiple testing procedure for controlling false discovery rates, *BMC Bioinformatics*, 18(1), pp. 1–10.

Love, M.I., Huber, W. and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2, *Genome Biology*, 15(12), p. 550.

McClintick, J.N. et al. (2014) Ethanol treatment of lymphoblastoid cell lines from alcoholics and non-alcoholics causes many subtle changes in gene expression, *Alcohol*, 48(6), pp. 603–610.

Menyhart, O., Weltz, B. and Gyorffy, B. (2021) Multipletesting.com: A tool for life science researchers for multiple hypothesis testing correction, *PLOS ONE*, 16(6), pp. 1–12.

Risch, N. and Merikangas, K. (1996) The future of genetic studies of complex human diseases, *Science*, 273(5281), pp. 1516–1517.

Storey, J.D. and Tibshirani, R. (2003) Statistical Significance for Genome-Wide Studies.

Tukey, J.W. (1977) Some thoughts on clinical trials, especially problems of multiplicity. *Science*, 198(4318), pp. 679–684.